Improving the Effectiveness of Content Popularity Prediction Methods using Time Series Trends

Flavio Figueiredo Marcos Gonçalves Jussara M. Almeida
Computer Science Department - Universidade Federal de Minas Gerais, Brazil
{flaviov, mgoncalv, jussara}@dcc.ufmg.br

ABSTRACT

We here present a simple and effective model to predict the popularity of web content. Our solution, which is the winner of two of the three tasks of the ECML/PKDD 2014 Predictive Analytics Challenge, aims at predicting user engagement metrics, such as number of visits and social network engagement, that a web page will achieve 48 hours after its upload, using only information available in the first hour after upload. Our model is based on two steps. We first use time series clustering techniques to extract common temporal trends of content popularity. Next, we use linear regression models, exploiting as predictors both content features (e.g., numbers of visits and mentions on online social networks) and metrics that capture the distance between the popularity time series to the trends extracted in the first step. We discuss why this model is effective and show its gains over state of the art alternatives.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms

Algorithms; Measurement

Keywords

analytics; predictive; challenge; web page; popularity; host

1. INTRODUCTION

With the ever-growing production of online content, characterizing and predicting user engagement (e.g., number of visits or social engagement such as Facebook likes) on web content may have multiple beneficial values such as: (1) understanding the human dynamics of information consumption; (2) supporting the decisions of content producers and providers on different tasks (e.g., marketing and content filtering); and, (3) understanding the physical processes that govern the growth of viewership on the Web. Several previous studies [2–4, 11] have characterized some of the factors that cause the popularity growth of different kinds of web content. Complementarily, various others [1, 6, 7, 11] have focused on the the task of popularity prediction. We focus here on the latter task, aiming at predicting the popularity of a piece of content.

Popularity prediction is a difficult and important task since it mostly translates into income and profits for content providers, creators and consumers alike. For example, more visitors to a web page may lead to more ad-clicks and sales. Moreover, content provisioning to a large amount of users may require decisions such as geographical sharding of content to servers (due to the increased traffic). Thus, if planning is not performed correctly, longer latencies and loading times, and thus, fewer users may be expected. Finally, accurate and early predictions can lead to better services to the end consumer, such as search engine rankings [8].

We here present a simple, and yet effective, model for predicting the popularity of online content. More specifically, we present the winning model of two of the three tasks of the ECML/PKDD 2014 Predictive Analytics Challenge. In the challenge, different features related to the popularity of 30,000 web pages from 100 different hosts were provided. The goal of the challenge is to predict the popularity of 30,000 other pages from the same 100 hosts 48 hours after their upload. The features provided for the task were measured in the first hour after upload for each page.

Our model exploits the temporal features related to web pages (e.g., past visits and social engagement), as well as typical popularity (i.e., number of visits) time series trends which exist in the dataset. Such trends are extracted via unsupervised learning methods. Specifically, it combines the temporal features with features that capture the distances between the popularity time series for each web page and the extracted trends. We present a data characterization that motivates the design of our solution, and show the gains in prediction accuracy (ranging from 15% to 27%) when it is compared to state of the art alternatives.

The rest of this paper is organized as follows. We formally describe the prediction problem and present the state of the art baseline methods in Section 2. In Section 3 we introduce our proposed solution, whereas our experiments and results are presented in Section 4. Finally, Section 5 concludes the paper.

2. BACKGROUND

We start this section by defining the content popularity prediction problem (Section 2.1). In this definition, as throughout the rest of the paper, we refer to a particular piece of content as a web page[1]. Next, we discuss existing state of the art solutions used as baselines in our experimental study (Section 2.2).

[1]We here focus on web page popularity prediction, given the goal of the ECML/PKDD Challenge. However, our models are general and can be applied to other types of online content.

ECML/PKDD Discovery Challenge on Predictive Analytics
Nancy, France, September 2014.